Overview of Existing Methods of Spam Mining and Potential Usefulness of Sender's Name as Parameter for Fuzzy String Matching

Soma Halder, Chun Wei, Alan Sprague

Department of Computer and Information Sciences
University of Alabama at Birmingham
Birmingham,Alabama,USA
{soma,weic,sprague}@cis.uab.edu

***Abstract*—This paper gives an overview and analyzes the different existing anti spamming techniques both content based, non content based methods and a combination of the duo. It also suggests that a small change in one of the parameters in the fuzzy string matching method could be useful to produce better results.**

***Index Terms* —email, spam, computer forensics, fuzzy string match.**

I. Introduction

The bulk of unsolicited email that floods our mail boxes is what we call spam today. Ever since its first appearance which happened about 30 years back in 1978 when Gary Theurk sent out the first spam email , the rising number of junk emails has been a great nuisance to the receivers of the email[1][2]. Today spam emails are far worse than the simple word 'junk' suggests because they bring terror in disguise trying to rob people's credentials like vital personal information, redirect recipients to phishing sites by making them click on malicious urls and above all spread malwares and viruses. Some of the malwares set up the recipient computer as a bot which in turn starts circulating spam emails. The July 2010 Symantec report says that 88.32 percent of emails were spam and 12% of these spam emails were used to spread malware [3].

This paper is broadly divided into three main parts : firstly we do a study on the background of spamming practices prevalent these days , secondly we do a brief analysis of the dominant anti spamming methods that have been developed so far and third we propose a method for mining spam and thus preventing their circulation.

II.BACKGROUND

In the early days of spamming, spammers used to send spam email directly to the recipients from their own mail address via the Internet Service Provider (ISP) mail server. The first spam email that went out in 3rd May 1978 to about 400 recipients of the ARPANET (what internet was then called), was sent by a marketer of Digital Equipment Corporation (DEC)[4] . However with the introduction of cyber crime laws and several anti-spam mechanisms, the spammers have been forced to be in disguise. Today most of the spam emails are sent through Botnets, Spamhauses, Fast Flux Service Networks (FFSN) and open relay sink holes in order to trespass the protective mechanism laid down by the spam filters, domain and IP black listings.

A. Definitions

Before we move into analyzing the different anti spam mechanisms we formally define the different methods taken up by the spammers to combat them and maintain their anonymity. Botnets are malware infected machines that are used for originating spam and spreading them to different recipients all over the web. If the recipient of the email gets infected by this malware they in turn act as bots. All bots are controlled by the Command and Control Server which is maintained by spammers [5].Backtracking a spam email with WHOIS information at any time would lead to a compromised machine whose owner is probably unaware that his machine is a bot.Spamhauses are fraud companies with domain names that last for a day and have been set up solely to lure investors to phishing sites. Fast Flux Service Networks are similar to the botnets, they make use of compromised machines that are used to host illegal websites[7].Open relay sinkholes are mail transfer agents(MTA) that are widely used by spammers to forward mail any sender or any recipient and they go undetected[5][6].

III.Anti Spamming Techniques

Anti Spamming methodsgenerally comprise (but not limited to) of two methods:

(a) Content Based Techniques.
(b) Non Content Based Techniques

Content based filtering refers to the filtering mechanisms that use the text portion in the spam emails. Whereas non content based techniques usually refer to methods that use DNS blacklisting. These days there are methods that use a flavor of both content and non-content based techniques in the same method (C. Wei *et al*).

A.. Content Based Techniques

Naïve Bayesian Filters

Bayesian filters use the Bayes' Theorem as a statistical method to detect spam. A training data set that contains

DOI: 01.IJIT.01.01.80

probabilities of the most recurring words in spam is created by feeding spam emails to the system. Then when test data gets fed to this system it finds the probability of the email to be spam or legitimate [8][9] based on Bayes' Theorem which is as follows:

the email to be spam or legitimate [8][9] based on Bayes' Theorem which is as follows:

$$P(S|W) = \frac{P(W|S).P(S)}{P(W|S).P(S)+P(W|L).P} \qquad (1)$$

where $P(S|\backslash$ is the probability of an Email to be Spam (S) when given a particular word (W)which is generally a part of spam email.

P(S) is the proportion of emails that are spam, P(W|S) is the probability of the word to appear in a spam message. Similarly P(W|L) is the probability of the word to appear in a legitimate message and P(L) is the proportion of emails that are legitimate . Putting the different probability values in the equation (1), we decide whether the test data(email) is spam or not.

Support Vector Machines (SVM)

SVM is a method of separating spam emails with *n* dimensional hyper planes. The most perfect hyper plane is called the decision boundary [12]. This decision boundary is used to separate the points in the first class from that in the second class[13].Often the points in the two classes are distributed and it is difficult to get a perfect hyperspace ,in such cases *n* dimensional hyperspaces are considered. The support vectors are used to define these hyperplanes. The training data and test data are mainly based on Term Frequency (TF) from spam messages [14].

Neural Networks

Neural Networks (NN) is another method of content based spam classification. Since neural networks have the capacity of working with large amount of data, they turn out to be very advantageous. Information in the neural network is generally in the form of numerical weights and connections. NN like other methods (overviewed earlier) first treats itself with a set of training data consisting of both spam and nonspam emails where phrases and words are the input vectors. Then when the system is fed with test data the NN finds a pattern and classifies messages as spam and non spam. Since NN weights are difficult to understand, visualization methods like interactive weight visualization have been introduced [22].

Random Forests and Active Learning

This method mainly uses term frequency (TF) and inverse document frequency (IDF) of email messages. This is followed by clustering to get the medoids. Medoids help in forming a forest. Active learning of the forest makes it more and more accurate. The authors of [15] report that accuracy is approximately 95% which might be only 65%, 66% for Naïve Bayes and SVM respectively. This is a very efficient method amongst all content based methods.

Content based methods have been facing serious hindrance from text obfuscation methods (Example: VIAGRA spelled as V1AGRA or as V-I-A-G-R-A) of late, but deobfuscation of content by CMOScript [10] or with the Hidden Markov Model [11] has brought up the spam hit rate.

B. Non Content Based Techniques

Domain Black Listing

Domain Blacklisting is a non content based method spam classification method. It is also known as real time black hole list(RBL).This method involves listing down all domain that ever appeared in the spam messages. SURBL and URIBL are two websites that maintains a database of these black listed domains[17][16].DNS lookups are used to query these databases[21].DNSBL is another such tool which keeps a track of all IP address that appeared through the domain name server in unsolicited emails[18]. However blacklists differ in their aggressiveness because some lists prefer giving less priority to false positives [21]. Domain blacklisting has a major disadvantage because spammers prefer to register a large number of new domain everyday (as domain hosting is pretty cheap) and get rid of them the next day. This evades detection by law enforcement as well as serves their purpose of spamming.

C. Combination of Content and Non-Content based methods

Domain Clustering using Fuzzy string matching Algorithm

This method proposed by C. Wei *et al* [19] involves using both the content of the email because it takes subject as a feature and also uses the IP address of the URL as another feature. The goal of this method is however much different from the other anti spamming methods discussed above because the input of this method consist of only spam emails unlike others. Once it has all the spam emails in hand it does a fuzzy matching of the both subject and IP. This is not a filtering method instead it clusters spam domains based on subject-IP pair similarity of the emails. The similarity is done by a fuzzy matching [19].

i. Subject-Matching

Spam emails are mostly generated using templates. Hence emails generated using same templates for some spam campaign are very similar. The method in [19] uses Inverse Levenshtein Distance (ILD) to do the fuzzy matching between subjects. The method involves the use of dynamic programming for finding the string alignment similarity. It is preferred the string similarity score have a value between 0 and 1 so they use Kulczynski Coefficient (defined in section IV) to get the score from the ILD value.

ii. Internet Protocol (IP) Address Matching

Since a single domain can be associated to several IP addresses, hence while matching two domains it is required to match all IP addresses associated with one domain to that of the other using Fuzzy matches. The maximum number of match gets considered [23]. prefer giving less priority to false positives [21]. Domain blacklisting has a major disadvantage

DOI: 01.IJIT.01.01.80

because spammers prefer to register a large number of new domain everyday (as domain hosting is pretty cheap) and get rid of them the next day. This evades detection by law enforcement as well as serves their purpose of spamming.

Two domains (or vertices) are connected to each other by a subject score which act as the edge between them. All domains that belong to same spamming campaign thus get connected by subjects. Connected components with similar strength (but they have to be above a threshold value) fall in the same cluster. This method successfully detected the Canadian Pharmacy spam that was prevalent in January 2010.

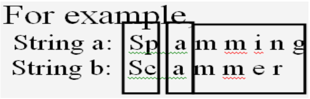
IV. Our methodology

A typical spam email (fig 1) shows that apart from the subject there is also a feature called the sender name. In our experiments we use the same methodology as used by [19][20] but we feed sender's name as a feature instead of subject.

. *Measuring similarity of Spam Sender's name*

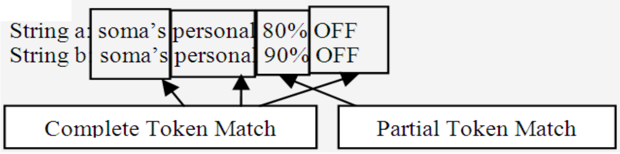
As discussed by [19] Inverse Levenshtein helps to finds alignment between strings.If **a** and **b** are two strings the ILD count for them would be as follows:

Strings **a**,**b** =InvLevDis(**a**,**b**).



InvLevDis(a,b)=4(as there are 4 matches)

We use the ILD to do a Fuzzy matching between Sender's names but often it is seen that the sender name is a sentence rather than a word separated by spaces in such cases we break it up in tokens and do ILD[20][23] .Token here represents each word in the sentence separated by spaces.

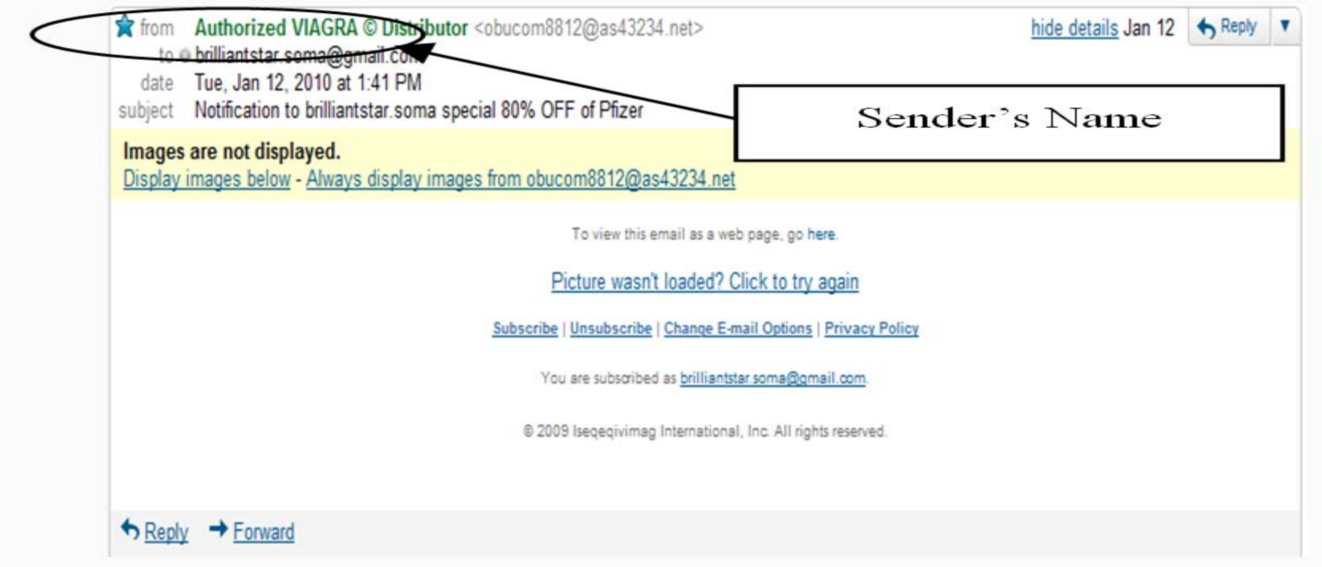

String a: soma's personal 80% OFF
String b: soma's personal 90% OFF
Complete Token Match
Partial Token Match

As discussed earlier by [19] that it preferable to have similarity score between 0 and 1 so apply Kulczynski Coefficient.

Kulczynski(**A,B**)=(ILD(A,B)|A| + ILD(A,B)/|B|)/2

Where |A|, |B| are the size of two strings.(i.e. the Senders name). ILD (A, B) is always nonnegative, and does not exceed min (|A|, |B|).

The rest of the procedure is the same as C. Wei *et al* where we find IP similarity and then cluster domains using connected components.



from Authorized VIAGRA © Distributor <obucom8812@as43234.net>

hide details Jan 12

to brilliantstar.soma@gmail.com

date Tue, Jan 12, 2010 at 1:41 PM

subject Notification to brilliantstar.soma special 80% OFF of Pfizer

Images are not displayed.

Display images below - Always display images from obucom8812@as43234.net

To view this email as a web page, go here.

Picture wasn't loaded? Click to try again

Subscribe | Unsubscribe | Change E-mail Options | Privacy Policy

You are subscribed as brilliantstar.soma@gmail.com

© 2009 Iseqeqivimag International, Inc. All rights reserved.

Sender's Name

Figure 1: Sample Spam Email

V. Results

Our experiment is based on a data of 298,680 spam emails. A complete Token Match suggested the maximum number of times a particular sender name repeated itself was 57,773 times and a partial token match suggested that the maximum count was 99,198 times. That is the cluster formed as a result is very big as compared to the total number of emails. The graph below (Figure 2) shows a comparison between the results for complete token match.

The graph contains the set of 5 most repeated sender names:

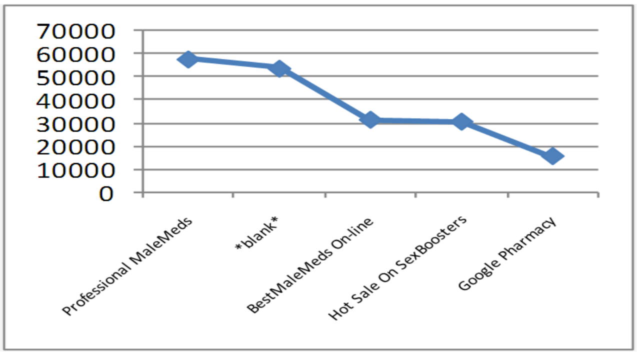


Figure 2:Graph for Complete Token Matching

Another experiment on a set of 312,585 spams separated into groups of 8 had the following result:

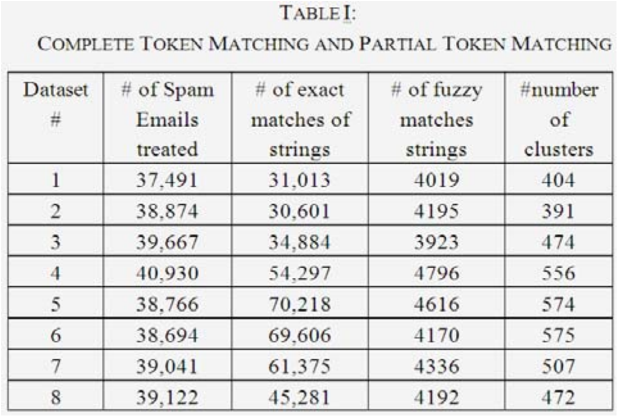


TABLE I:
COMPLETE TOKEN MATCHING AND PARTIAL TOKEN MATCHING

Dataset #	# of Spam Emails treated	# of exact matches of strings	# of fuzzy matches strings	#number of clusters
1	37,491	31,013	4019	404
2	38,874	30,601	4195	391
3	39,667	34,884	3923	474
4	40,930	54,297	4796	556
5	38,766	70,218	4616	574
6	38,694	69,606	4170	575
7	39,041	61,375	4336	507
8	39,122	45,281	4192	472

VI. CONCLUSION

Seeing the graph above (Figure: 2) we see that sender's name can definitely be considered as one of the parameters while doing fuzzy matching with strings. Table 1 justifies this argument by giving the total count of the number clusters formed as a result of fuzzy match. There by we suggest that sender's name that appears in all spam emails definitely can be taken as a feature when clustering domains in the spam emails. Therefore the most obvious future work in this direction would be to cross validate C. Wei *et al*'s clusters with the clusters obtained by the methods of this paper by making them work on the same data set.

REFERENCES

[1] B. Templeton, "Origin of the term 'spam' to mean net abuse".Available:http://www.templetons.com/brad/spamterm.html.

[2] H.Cox, "The History of Spam emails" Available : http://ezinearticles.com/?The-History-of-Spam-Emails&id=785433

[3] Symantec, *State of Spam and Phishing, A monthly Report* ,Report #43,July 2010.

[4] B. Templeton, "Reaction to the DEC Spam of 1978" . Available : http://www.templetons.com/brad/spamreact.html.

[5] A. Pathak, F. Qian, Y. C. Hu, Z. M. Mao, and S. Ranjan, "Botnet Spam Campaigns Can Be Long Lasting:Evidence, Implications, and Analysis" .*In Proc. of the 11th Int.Joint Conf.on Measurement and Modeling of Computer System*, 2009.

[6] A. Pathak , Y. C. Hu, Z. M. Mao, "Peeking into Spammer Behavior from a Unique Vantage Point".*In Proc. of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threat*, 2008.

[7] T. Holz, C. Gorecki, K. Rieck, F. C. Freiling, "Measuring and Detecting Fast-Flux Service Networks". *In Proc. Of 16th Annu. Network & Distributed System Security Symposium*, 2008.

[8] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz, "A Baysian Approach to Filtering Junk Email".In Proc. of AAAI-98 Workshop on Learning for Text Categorization., 1998.

[9] G. Robinson, "A statistical approach to the spam problem,*Linux Journal*,Issue # 10, pp 3 2003.

[10] CMOScript, Available : http://sandgnat.com/cmos

[11] H. Lee, A. Y. Ng, "Spam Deobfuscation using a Hidden Markov Model". *In Proc. of the 2nd Conf. on Email and Anti-Spam*,CEAS,2005.

[12] C. Liu, "Experiments on Spam detection with Boosting, SVM and Naïve Bayes", *UCSC.*

[13] A. Khorsi, "An Overview of Content Based Spam Filtering Techniques". *Informatics 31* pp 269-277, 2007.

[14] H. Drucker, Donghui, V. N. Vapnik. "Support Vector Machines for Spam Categorization". *AT&T Labs-Research,*1999.

[15] D. Debarr,H. Wechsler, "Spam Detection using Clustering,Random Forests and Active Learning". *In Proc. of the 6th Conf. on Email and Anti-Spam* ,CEAS , 2009.

[16] Available : http://www.surbl.org/.

[17] Available : http://www.uribl.com/.

[18] Available : http://www.dnsbl.info/.

[19] C. Wei,A. Sprague,G. Warner,A. Skjellum, "Identifying New Spam Domains by Hosting IPs: Improving Domain Blacklisting". *In Proc. of the 7th Conf. on Email and Anti-Spam* ,CEAS,2010.

[20] C. Wei, "A malware-generated Spam Emails with a Novel Fuzzy String Matching Algorithm". *In Proc. of the ACM symposium on Applied Computing*, 2009.

[21] D. Cook, J. Hartnett, K. Manderson and J. Scanlan. "Catching Spam Before it Arrives: Domain Specific Dynamic Blacklists". *In Proc. of the 2006 Australasian workshops on Grid computing and e-research,* 2006.

[22] Fan-Yin Tzeng,Kwan-Liu Ma(2005).Opening the Black Box :Data Driven Visualization of Neural Networks. *Visualization, 2005. VIS 05. IEEE*.Page 383-390.

[23] C. Wei,A. Sprague,G. Warner, "A. Skjellum.Clustering Spam Domains and Destination websites :Digital Forensics for Data mining".*Journal of Digital Forensics,Security and Law,(2009)*